KORPUSLINGUISTIK – DAS UNBEKANNTE WESEN
oder
Mythen über Korpora und Korpuslinguistik

von Rainer Perkuhn und Cyril Belica

Sind Korpora nur Belegsammlungen oder Zettelkästen in elektronischer Form? Mitnichten! In entsprechender Größe (vgl. Church / Mercer 1993) und mit den entsprechenden Analysemethoden eröffnen sie eine eigene Perspektive in der linguistischen Forschung – die korpuslinguistische Perspektive.

> „In a corpus-driven approach the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence. The corpus, therefore, is seen as more than a repository of examples to back pre-existing theories or a probabilistic extension to an already well defined system. The theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus. Indeed, many of the statements are of a kind that are not usually accessible by any other means than the inspection of corpus evidence. Examples are normally taken verbatim, in other words they are not adjusted in any way to fit the predefined categories of the analyst; recurrent patterns and frequency distributions are expected to form the basic evidence for linguistic categories; the absence of a pattern is considered potentially meaningful."
> (Tognini-Bonelli 2001, S. 84)

Aus dieser Perspektive spielen einzelne Belege nur insofern eine Rolle, als dass sie ihren – normalerweise winzigen und fast vernachlässigbaren – Beitrag zu einem Gesamtbild leisten. Die korpuslinguistische Herausforderung besteht darin, Linguisten Möglichkeiten an die Hand zu geben, dieses Gesamtbild zu „erkennen", wobei einerseits die Körnung des Bildes nicht zu grob sein darf, andererseits die Konturen auf dem Bild deutlich genug sein müssen, um es interpretieren zu können. Beide Punkte setzen Sprachdatenmassen in ausreichendem Umfang und geeignete Analysemethoden voraus. Das Anliegen der Korpuslinguistik ist – wohlgemerkt – die Analyse und Beschreibung des Sprachgebrauchs, normative Aussagen sollten stets losgelöst davon formuliert und gedeutet werden. Das IDS, dessen Ziel satzungsgemäß die „Erforschung und Dokumentation der deutschen Sprache in ihrem gegenwärtigen Gebrauch" ist, hat die Forschung in diesem Paradigma organisatorisch in einem neu eingerichteten Programmbereich „Korpuslinguistik" verankert.

Eine angemessene, sachgemäße Diskussion über Stärken und Schwächen, Möglichkeiten und Grenzen der Korpuslinguistik ist überschattet von vielen Mythen, die sich mittlerweile eingebürgert haben und die in vielen Diskussionen – gerade unter Linguisten – immer wieder aufkommen. An dieser Stelle möchten wir einige der verbreitetsten Mythen zusammenstellen und die Hintergründe aus dieser korpuslinguistischen Perspektive erörtern (vgl. Perkuhn et al. 2006).

IMPRESSUM

Herausgeber: Institut für Deutsche Sprache, Postfach 101621, 68016 Mannheim.
Internet: http://www.ids-mannheim.de
Mitglied der Leibniz Gemeinschaft

Redaktion: Annette Trabold (Leitung), Karl-Heinz Bausch, Heidrun Kämper, Horst Schwinn, Eva Teubert
Redaktionsassistenz: Jens Gerdes, Anne Steinz
E-Mail: sprachreport@ids-mannheim.de

Satz & Layout: Claus Hoffmann (IDS)
Belichtung: Afosatz Frey, 68199 Mannheim
Druck: Morawek, 68199 Mannheim
gedruckt auf 100% chlorfrei gebleichtem Papier
ISSN 0178-644X

Auflage: 2500, Erscheinungsweise: vierteljährlich
Jahresabonnement: 10,– EUR Einzelheft: 3,– EUR
Bezugsadresse: Institut für Deutsche Sprache, Postfach 10 16 21, 68016 Mannheim
Tel. 0621/1581-0

In eigener Sache – an die Autoren:

Wir bitten Sie, Ihre Beiträge als WINWORD oder RTF-Datei im Anhang per E-Mail zu schicken an:
sprachreport@ids-mannheim.de oder auf Diskette.

Bitte wählen Sie dazu folgendes Disketten-Format:
3.5 Zoll, WINDOWS-formatiert.

NICHT bearbeiten können wir:
– 5.25 Zoll-Disketten,
– MAC-formatierte Disketten.

Die Texte sollten nicht mit komplizierten Layouts und ohne Formatvorlage erstellt sein, die Formatvorlagen erstellen wir.

Der SPRACHREPORT wird mit PageMaker 6.5 erstellt.

Mythos 1:

Korpuslinguistische Methoden erfordern sprachliches / linguistisches (Hintergrund-)Wissen.

Dies stimmt nicht. Korpuslinguistische Methoden basieren auf denselben Prinzipien wie etwa die Warenkorbanalyse oder andere Data Mining-Verfahren[1]. Wenn die Warenkorbanalyse einen Hinweis auf einen Zusammenhang zwischen Nudeln, Ketchup und Hackfleisch liefert, muss das Verfahren nicht wissen, was Spaghetti Bolognese ist. Entsprechend liefern korpuslinguistische Methoden einen Hinweis auf Zusammenhänge zwischen Wörtern (und/oder Textmerkmalen) ohne sprachliches (Hintergrund-)Wissen (etwa über Syntax oder Valenzen). Dass man aus den Zutaten Spaghetti Bolognese zubereiten kann, ist ein Vorgang einer anderen Qualität, der losgelöst vom Einkaufen verstanden werden kann und der auch eine andere Kompetenz (nämlich die des Kochens) erfordert. Analog schließt sich an die korpuslinguistische Analyse eine sprachliche und/oder linguistische Interpretation an, die ebenfalls eine andere Kompetenz voraussetzt.

Eine besondere Herausforderung besteht bei diesen (Data Mining- und korpuslinguistischen) Verfahren darin, Muster von Zusammenhängen auf abstrakteren Ebenen zu erfassen. Analog zu dem o.g. Beispiel liefern die Analysen auch Zusammenhänge zwischen den Zutaten zu anderen Rezepten, aber auch Gruppen wie z.B. „Pampers, Babynahrung, Ohrstöpsel“ oder „Sekt, Chips, Pappbecher, Kopfschmerztabletten“, deren Zusammensetzung sicher ganz anders motiviert ist, sei es durch Nachwuchs oder das Organisieren einer Party. Auch bei der Analyse der Sprache werden Zusammenhänge aufgedeckt, für die Teildisziplinen der Linguistik Erklärungen anbieten können, die auch ein Korpuslinguist kennen und nicht außer Acht lassen sollte – für die meisten Zusammenhänge lassen sich aber (zumindest auch) andere Motivationen finden. Und besonders spannend wird es dann, wenn aus neuen Erkenntnissen über sprachliche Phänomene neue Ansätze hervorgehen, die bereits Bekanntes mit einfacheren Konzepten oder sogar bis dato unbekannte Phänomene erklären können.

Mythos 2:

Korpora gewinnen an Wert, wenn sie annotiert werden.

Die Korpora selber – als Abbild eines Ausschnitts der sprachlichen Realität – gewinnen durch die Annotationen nicht an Wert. Bei „berechenbaren“ Annotationen können Anfragen/Analysen evtl. schneller bearbeitet werden. Nutzt man bei Anfragen hingegen „interpretierte“ Annotationen, liefern die Ergebnisse lediglich ein Abbild der Qualität der Annotationen, nicht der empirischen Daten („Man kann nur die Ostereier finden, die man selber versteckt hat.“ – Erben 2003).

Es soll nicht in Abrede gestellt werden, dass es Szenarien für sinnvolle Anwendungen von annotierten Korpora gibt. Bibliografische Textannotationen zu Autor, zeitlicher und regionaler Herkunft oder thematischer Beschaffenheit erlauben etwa die Inbezugsetzung der Sprachdaten zu diesen Informationen. Für die meisten Annotationen anderer Art gilt allerdings – solange sie nicht ohne Vorannahmen aus den Daten selbst hervorgetreten sind –, dass das Korpus lediglich zu einer Testfallsammlung der Vorannahmen degeneriert.

Mythos 3:

Die Existenz eines schlechten Belegs beweist, dass es nicht sinnvoll ist, Belege anzugeben.

Grundsätzliches zu diesem Mythos: Die Korpuslinguistik interessiert sich nicht für Einzelbelege. Ein Einzelbeleg, d.h. eine einzelne Textstelle, ist lediglich ein Mosaikstein in einem Bild, das erst in seiner Gesamtheit erkennbar ist. Gelingt es, eine beliebige korpuslinguistische Aussage durch das Weglassen eines Belegs zu erschüttern, so handelt es sich definitionsgemäß ohnehin um keine valide korpuslinguistische Aussage. Daraus folgt unmittelbar, dass Aussagen, die sich jeweils nur auf ganz geringe Belegzahlen stützen, in ihrer Gesamtheit keine korpuslinguistischen Aussagen sind.

Vor dem Hintergrund einer lexikographischen Anwendung verfolgt man mit der Angabe von Belegen verschiedene Ziele. Hierbei gilt insbesondere, dass die Qualität eines Belegs nur vor dem Hintergrund eines Angabezwecks zu bewerten ist. Ein Beleg kann u.a.

- ein Einzelvorkommen dokumentieren
  … das kann er immer, die Qualität ist per se o.k.;
- eine prototypische Verwendungsweise illustrieren;
  … dann ist die Qualität vom Auswahlproblem determiniert;
- definitorische Funktion suggerieren
  … inwieweit dies sinnvoll ist, hängt davon ab, ob man einem einmaligen Gebrauch eine definitorische Wirkung zuerkennen möchte; ist die „Bedeutung“ von A gleichzusetzen mit der von B, nur

weil bei einer Gelegenheit geäußert wurde „A sei B“?

1. Am besten sollte man solche Belege verwerfen;
2. sie sind evtl. sinnvoll, wenn es sich um „Neologismen“ handelt, zur Dokumentation ihrer Entstehung;
3. man sollte unterscheiden zwischen „mentioned, not used“; „mentioned“ ist normalerweise ungeeignet als Beleg, es sei denn, das ist der typische Gebrauch

… grundsätzlich ist dieser Anspruch für einen Beleg zu hinterfragen: Ein Beleg zeigt eine Verwendungsweise; daraus lässt sich weder schließen, dass das Wort nur so gebraucht werden kann, noch wie andere Verwendungsweisen zu handhaben sind („ein Schwein ferkelt“ lässt sich durchaus dahingehend verstehen, dass ein Schwein Ferkel bekommt; dass dieser Vorgang weiblichen Tieren vorbehalten ist, lässt sich über das Weltwissen erschließen; der Beleg legt aber in keiner Weise nahe zu vermuten, dass man auch „das Wildschwein ferkelt“ oder „der Eber ferkelt“ sagen kann).

Statt auf Belege zu verzichten, wäre sicher die bessere Lösung, bei Bedarf intelligent Belegmengen dynamisch zusammenzustellen.

Mythos 4:

Man kann / darf / soll negative Aussagen aufgrund von Korpusbefunden formulieren.

Korpusbefunde können nur positiv interpretiert werden. Wenn ein Phänomen in einem Korpus belegt ist, dann existiert es auch. Wenn ein Phänomen in einem Korpus nicht belegt ist, kann man keine Aussage folgern. Man kann z.B. nicht sagen: „Eine Form bzw. Formulierung ist nicht belegt, deshalb ist sie nicht akzeptabel“. In diesem Fall kann es immer zwei Erklärungen geben: Entweder existiert das Phänomen tatsächlich nicht, oder das Korpus ist in der Hinsicht unvollständig. Man kann in einer Annäherung höchstens pseudo-quantitative Aussagen wagen: In den Korpora Häufiges ist auch in der Realität häufig, Seltenes ist selten, Nicht-Beobachtetes existiert nicht oder ist sehr wahrscheinlich selten – in Abhängigkeit von der Extrapolierbarkeit der Aussage über den zugrunde gelegten Sprachausschnitt. Je größer Korpora werden, desto mehr seltene Phänomene werden abgedeckt. Daraus, dass eine bestimmte Form nicht belegt ist, z.B. *Hühnerkäfigs*, darf nicht abgleitet werden, dass diese nie gebraucht wird oder nie gebraucht werden kann. Noch deutlicher sieht man das an „Wortkombinationen“: *Verdis Roman Guernica*. Auch wenn diese Kombinationen nicht in den Korpora belegt sind und auch vielleicht tatsächlich noch nie geäußert wurden, heißt das doch noch lange nicht, dass sie nicht verwendet werden können.

Mythos 5:

Internetsuchanfragen liefern bessere Ergebnisse als Korpusanalysen.
Wenn die Internetsuchmaschine „Google“ viele Treffer für Suchobjekte vermeldet, müssten diese Suchobjekte auch bei Korpusanalysen hervortreten.

Dieser Vergleich hinkt aufgrund der unterschiedlichen (Qualität der) Datenbasis und des Verfahrens. Internetsuchanfragen haben sicher auch für linguistische Untersuchungen einen großen Wert. Das Internet stellt ein riesiges Reservoir an Information zur Verfügung. Es ist u.E. aber vollkommen unklar, welcher Typ / welcher Ausschnitt der Sprache sich im Internet manifestiert. Selbst wenn Korpora zeitungslastig sind, sollte man im Hinterkopf haben, dass das Internet „elektronische Medien“-lastig ist. Beide Ausschnitte müssen kritisch hinterfragt werden, wenn sie dazu genutzt werden, Schlüsse über den allgemeinen Sprachgebrauch zu ziehen.

Grundsätzlich haben Internetsuchanfragen eine ganz spezielle Funktion. Mit Hilfe der Angabe von Schlüsselwörtern sollen Dokumente im Internet aufgespürt werden. Gibt man mehrere Suchwörter ein, werden alle Dokumente als Treffer betrachtet, in denen diese Suchwörter an irgendeiner Stelle vorkommen. Ein Zusammenhang zwischen diesen Suchwörtern als Begriffen ist eventuell durch das in dem Dokument beschriebene Thema begründet, vielleicht aber auch ganz zufällig. Alternative Phrasensuchen erlauben die zusätzliche Forderung, dass die Suchwörter unmittelbar aufeinander folgen müssen. Dies mag der damit verbundenen Intention näher kommen, leidet aber genauso darunter, dass es eben eine Suchanfrage ist und Verhältnismäßigkeiten zwischen den Vorkommen der beteiligten Suchobjekte außer Acht lässt. Wenn z.B. die Phrasen *Reise unternehmen* und *Reise machen* oft im Internet dokumentiert sind, sagt dies noch nicht viel über den inneren Zusammenhang. Eine Suchanfrage an ein Korpus könnte auch viele Treffer belegen (sogar mit vorgegebenen maximalen Wort- oder Satzabständen). Im Gegensatz zu einer Suche ist der Zweck einer Korpusanalyse zu hinterfragen: Ist das häufige gemeinsame Vorkommen rein zufälliger Natur oder durch eines der beteiligten Wörter bedingt? Dass z.B. in der Nähe des

Wortes *ist* das Wort *der* häufig vorkommt (eine Anfrage an „Google“ liefert viele Treffer), ist nicht verwunderlich, weil das Wort *der* sehr häufig in der deutschen Sprache ist (das Wort *ist* ist natürlich auch sehr häufig). Beliebige Kombinationen von einem beliebigen Wort und dem Wort *der* werden oft vorkommen, allein deshalb, weil das Wort *der* oft vorkommt. Die Kombination der Wörter *Zähne* und *machen* ist sicher sowohl im Internet als auch in den Korpora sehr häufig, die Kombination der Wörter *Zähne* und *putzen* sehr viel seltener belegt. Da aber die beteiligten Wörter im Vergleich zu *machen* noch sehr viel seltener sind, ist die letztere Kombination viel auffälliger. Das obige Beispiel *Reise machen* ist deshalb weniger auffällig, da *machen* viel häufiger ist und in der Nähe von vielen anderen Wörtern auch häufig vorkommt. Rechtfertigt dies, dass auch sämtliche anderen Kombinationen mit *machen* berücksichtigt werden? *Reise* ist 84.143 mal im Deutschen Referenzkorpus des IDS (DEREKO 2005) belegt, *machen* 633.293 mal; diese riesige Anzahl führt aber nur zu 1.248 Belegen für *Reise* und *machen* innerhalb eines Satzes. Das Wort *unternehmen* ist (im Vergleich zu *machen*) nur 20.477 mal belegt, die Kombination *Reise* und *unternehmen* innerhalb eines Satzes jedoch 404 mal. Dieser Wert ist im Verhältnis zu dem, den man erwarten dürfte, sehr viel höher als der Wert bei *Reise machen*. Deshalb ist es legitim, wenn die Analyse *Reise unternehmen* als signifikante Verbindung hervorbringt, *Reise machen* aber nicht. Um die „Auffälligkeit“ beurteilen zu können, muss man die Gesamthäufigkeiten kennen (und nicht nur erahnen!). Das ist im Internet – anders als im Korpus – nicht (leicht oder zuverlässig) möglich. Die Ergebnisse von Internetsuchanfragen werden noch lange unter der mangelnden Genauigkeit leiden, mit der die Treffermenge das mit der Suchanfrage intendierte Phänomen trifft („precision“). Anders verhält es sich, wenn Daten aus dem Internet als Korpora aufbereitet werden. Damit kann man z.T. zumindest die Breite der Daten aus dem Internet einfangen, erlaubt aber auch gleichzeitig die Anfrage- und Analysemöglichkeiten, wie man sie z.B. von COSMAS (COSMAS II 2005) gewohnt ist. Erst dadurch werden zwei Typen von Suchanfragen möglich: 1) „Gibt es ...?“ oder „Zeig‘ mir ...!“ als Beleg für die Existenz eines Phänomens, wie es über Suchmaschinen abgefragt werden kann, aber auch 2) „Wie typisch ...“ oder „Was sollte ich über ... sagen?“ als Evidenz für Typisches und Auffälliges.

Der fundamentale Unterschied zwischen einer Suche und einer Analyse besteht aber darin, dass man bei der Suche vorher wissen muss, wonach man suchen möchte. Die Frage ist, ob eine Suche nach *Zähne* und *putzen* tatsächlich jedem in den Sinn käme. Die Analyse bringt diese Verbindung hervor, ohne dass der Linguist sie erahnen müsste.

Mythos 6:

Kleine Korpora sind besser als große.

Für spezielle Anwendungen, insbesondere solcher Art, die auf die Auswahl weniger Belege zielen, mag dies stimmen. Kleine Korpora sind in vielerlei Hinsicht leichter zu handhaben als große, insbesondere, wenn sie für spezifische Problemstellungen aufgebaut und gepflegt werden. Sie bergen aber natürlich das Risiko, dass allgemeinere, über die der Korpuskomposition zugrunde liegende Problemstellung hinaus gehende Suchanfragen häufig zu Fehlanzeigen bzw. Analysen zu minderwertigen Ergebnissen führen. Große Korpora verbessern die Chancen, „gute“ Belege zu finden.[2] Dies ist allerdings nicht Gegenstand der korpuslinguistischen Methoden. Je größer die Datensammlungen sind, desto mehr seltene Phänomene decken sie mit ausreichender Aussagekraft ab. Für speziellere Fragestellungen lassen sich aus den grundsätzlich zur Verfügung stehenden Daten (Archiven) kleine „virtuelle“ Korpora definieren. Aber erst ab einer gewissen Größenordnung enthalten Korpora mehr Wissen als die Summe ihrer Belege. Auf dieses latente Wissen zielen die korpusanalytischen Methoden, für deren Anwendbarkeit eine kritische Datenmasse mindestens zur Verfügung stehen muss.

Mythos 7:

Korpora enthalten manchmal „Quatsch“.

Eine technisch einwandfreie und authentische Abbildung vorausgesetzt, enthalten Korpora nur Fakten über den Sprachgebrauch. Sie erfassen grammatisch korrekten und falschen Umgang mit der Sprache, so wie ihn die abgebildete Sprachgemeinschaft praktiziert. Darüber hinaus findet sich sicher auch kreativer Umgang mit der Sprache, der von Grammatikschreibern nicht antizipiert werden kann. Eventuell wird – vielleicht sogar bewusst – in Kauf genommen, dass Formulierungen grammatisch nicht zulässig sind, wie z.B. „Ick liebe dir“. Es steht den korpusdateninterpretierenden Linguisten natürlich frei, die Daten entsprechend zu kategorisieren. Als „Quatsch“ können die Daten nur in Hinsicht auf eine bestimmte Anwendung und somit als Folge einer Interpretation gewertet werden. Für jedes Datum, das aus einer Sicht als „Quatsch“ eingeordnet wurde, findet sich stets eine andere Sicht,

für die das Datum sinnvoll interpretiert werden kann. Ein Japaner z.B., der Deutsch als Fremdsprache gut beherrscht, sogar die korrekte grammatische Form kennt, möchte vielleicht trotzdem wissen, dass in einem bestimmten Kontext „Ick liebe dir“ verwendet wird, um auf Berliner Dialekt anzuspielen.

Mythos 8:

Korpora sagen einem, wie gesprochen wird bzw. gesprochen werden soll.

Grundsätzlich können Korpusdaten nur die Sprache der Vergangenheit (bis an die Gegenwart[3] heran) erfassen. Aus diesen Daten lassen sich quantitative Aussagen ableiten, aber nur schwerlich qualitative über Akzeptabilität oder Wohlgeformtheit. Ein formal grammatisches Urteil kann lediglich über eine Interpretation zu den Daten hinzugefügt werden. Eine „pragmatisch-evolutionäre“ Grammatikalitätsbeurteilung findet sich aber schon in einer gewissen Weise in den Daten versteckt: In der Evolution der Sprache werden Formulierungen, die eine Mehrheit nicht akzeptabel findet, „besseren“ Formulierungen unterlegen sein und sich somit nicht als normale oder typische Formulierung (für eine bestimmte Zeit oder für einen bestimmten Raum) durchsetzen. Diese Typikalität ist wiederum etwas, was sich quantitativ ermitteln lässt und eines der – wenn nicht sogar das – Hauptanliegen korpuslinguistischer Methoden.

Mythos 9:

Ein Korpus ist schlecht, wenn ein Forscher im Korpus nicht das findet, wonach er sucht.

Wenn ein Forscher bereits so gezielt suchen kann, dass er auch entscheiden kann, ob er etwas Passendes findet oder nicht, dann braucht er kein Korpus mehr – die von ihm gestellte Frage hat er bereits beantwortet, korpuslinguistische Verfahren helfen ihm in diesem Fall nicht weiter. Für die Bestätigung seiner „Antwort“ sind Korpora genauso gut oder sogar eher schlechter geeignet, als andere empirische Quellen wie z.B. die bereits diskutierte Internetanfrage oder gezielte Feldstudien.

Mythos 10:

Zu einem Erkenntnisgewinn können einem Forscher Korpora nur dann verhelfen, wenn sie vor dem Hintergrund eines bereits bekannten linguistischen Modells oder einer bekannten linguistischen Theorie ausgewertet werden.

Dies ist im Prinzip das Kondensat aller bisher diskutierten Mythen. Wie wir bereits in der Erläuterung zu Mythos 1 angedeutet haben, verzichten korpuslinguistische Verfahren auf Annahmen bezüglich linguistischer Modelle oder Theorien – und liefern trotzdem z.T. verblüffende Erkenntnisse. Damit wollen wir nicht sagen, dass wir sämtliche Modelle und Theorien über Bord werfen sollen. So ganz verkehrt kann nicht sein, was so viele Forscher in so langer Tradition erarbeitet haben. Nur war diesen Forschern der Blick auf so viel Sprache auf einmal nicht möglich, wie es korpuslinguistische Verfahren erlauben. Ein Verzicht auf traditionelle Herangehensweisen eröffnet die Möglichkeit, zunächst die Sprache für sich selbst sprechen zu lassen – und dann zu schauen, inwieweit die hervorgetretenen Phänomene sich mit dem klassischen linguistischen Denkapparat erklären lassen.

Kategorien können hilfreich sein, um die Erkenntnisse zum Ausdruck zu bringen; ihre Existenz vorauszusetzen kann aber auch manchmal den Blick für das Wesentliche verschließen. Anders verhält es sich, wenn sich eine Systematik von „Zusammengehörigem“ aus der Analyse und Interpretation der Daten ergibt:

> „Eine fundamentale Aufgabe jeder Wissenschaft ist die Schaffung einer Ordnung, das Finden von Mustern in der Menge mannigfaltiger, unübersichtlicher Daten. Klassifikations-, Korrelations-, Mustererkennungs- und andere induktiv-heuristische Verfahren dienen hauptsächlich dem Zweck, neue, zuvor nicht bekannte Phänomene und Zusammenhänge zu entdecken, zumal wenn, wie in der Korpuslinguistik, die Daten wegen ihrer schieren Masse mit dem Intellekt nicht einmal gesichtet werden könnten. Tatsächlich beruhen viele Erkenntnisse auf empirischen Generalisierungen, die nachträglich deduktiv verankert [...] wurden.“ (Köhler 2005, S. 4f)

Versuch, eine korpuslinguistische Leitlinie zu skizzieren

In der obigen Diskussion verschiedener Mythen haben wir vor allem dargestellt, wie wir Korpuslinguistik nicht verstanden wissen möchten, aber nur versteckt angedeutet, wie diese positiv über Eigenschaften beschrieben werden kann. Dies wollen wir nun ansatzweise mit einer Leitlinie andeuten, mit deren Hilfe wir hinterfragen wollen, inwieweit

korpusbasierte Ansätze diese Maxime tatsächlich zur Grundlage ihres Arbeitens und Denkens machen (wie z.B. in Belica / Steyer 2006).

Die Doktrin (streng) korpuslinguistischen Denkens ergibt sich schrittweise aus einer täglich bei jedem Kleinkind beobachtbaren Feststellung:

- Alles, was man wissen muss, um eine Sprache zu erwerben, steckt in der Sprache selbst.[4]
- Alles, was man wissen muss, um eine Sprache zu erlernen, steckt in der Sprache selbst.
- Alles, was man wissen muss, um eine Sprache zu verstehen, steckt in der Sprache selbst.
- Alles, was man wissen muss, um eine Sprache zu vermitteln, steckt in der Sprache selbst.
- Alles, was man wissen muss, um Erkenntnisse über eine Sprache zu gewinnen, steckt in der Sprache selbst.
- Alles, was man wissen muss, um Erkenntnisse über eine Sprache zu gewinnen und zu vermitteln, steckt in der Sprache selbst.
- Alles, was man wissen muss, um Erkenntnisse über Sprache zu gewinnen, steckt in den Sprachen selbst.
- Alles, was man wissen muss, um Erkenntnisse über Sprache zu gewinnen und zu vermitteln, steckt in den Sprachen selbst.

Die ersten vier Schritte sind in der Fremdsprachendidaktik und dem Forschungsansatz des „data driven learning“ bereits auf fruchtbaren Boden gefallen.

> „Es ist in der Fremdsprachendidaktik unumstritten, dass das Sprachenlernen soweit wie möglich induktiv gestaltet werden sollte. Daher gilt es beispielsweise, in der Wortschatzvermittlung Schülerinnen und Schülern Strategien zu vermitteln, wie sie in authentischen Texten zunehmend selbstständig die Bedeutung unbekannter Wörter aus dem Kontext erschließen können, ohne ständig auf deduktiv vorgegebene Wörterbuchdefinitionen zurückgreifen zu müssen [...].“ (Mukherjee 2002, S. 67)

Das „Alles, was man wissen muss“ entnimmt man nicht einzelnen Beispielen oder Fällen, sondern der Spracherwerbende / -lerner / Linguist entnimmt diese als Regelmäßigkeiten, Gesetzmäßigkeiten über Muster aus einem massenhaften Gebrauch von Sprache. Da außer dem Muttersprachler den anderen Interessierten aber die Zeit fehlt, um zu warten, bis sich ihnen die Muster von selber „auftun“, brauchen sie Unterstützung, um den Zeitfaktor auszugleichen. Sie brauchen zusammengefasst und komprimiert die Sammlung von Sprachdaten, denen Muttersprachler über einen langen Zeitraum ausgesetzt sind. An dieser Stelle setzen korpuslinguistische Methoden an und versuchen, Strukturen im massenhaften Gebrauch von Sprache aufzudecken bzw. für eine weitere Interpretation vorzubereiten (vgl. u.a. Steyer 2004).

> „Corpora provide no direct evidence for meanings. Meanings are inferred from contexts in reading texts in a corpus, in much the same way that meanings are inferred in reading any other kind of text, but with this difference: by seeing many uses of the target word in close proximity, the analyst can identify groups of normal uses of the target word according to their common syntagmatic features. A large corpus provides evidence of the patterns of usage with which meanings are associated. The larger the corpus, the more strikingly the patterns stand out.“ (Hanks 2004, S. 246)

Die Korpuslinguistik aus unserer Perspektive möchte zeigen, dass alle Ableitungsschritte bis zu der Doktrin „Alles, was man wissen muss, um Erkenntnisse über Sprache zu gewinnen und zu vermitteln, steckt in den Sprachen selbst“ ihre Berechtigung haben und dass ihre Verinnerlichung auch für andere linguistische Disziplinen einen Gewinn darstellt.

> „Analysis of extended naturally occurring texts, spoken and written, and, in particular, computer processing of texts have revealed quite unsuspected patterns of language [...] The big difference has been the availability of data [...]
>
> [The] major novelty was the recording of completely new evidence about how language is used [...]
>
> [The] contrast exposed between the impressions of language detail denoted by people, and the evidence compiled objectively from texts is huge and systematic [...]
>
> The language looks different when you look at a lot of it at once [...]“
>
> (Sinclair 1991, S. xvii, 1, 2, 4, 100)

Solange dieses Bestreben, sich nur am puren Sprachgebrauch zu orientieren (d.h. ohne Vorannahmen) und davon „vieles auf einmal betrachten zu wollen“, nicht erkennbar ist, kann man u.E. nicht von einem korpuslinguistischen Vorgehen i.e.S. sprechen.

Literatur

Belica, Cyril / Steyer, Kathrin (2006): Korpusanalytische Zugänge zu sprachlichem Usus. In: AUC (Acta Universitatis Carolinae), GERMANISTICA PRAGENSIA XX. Praha: Karolinum, erscheint 2006 (Vorabdruck als pdf unter www.ids-mannheim.de/lexik/UsuelleWortverbindungen/CBKSpraha.ver20050426. mit.summ.pdf).

Church, Kenneth W. / Mercer, Robert L. (1993): Introduction to the Special Issue on Computational Linguistics Using Large Corpora. Computational Linguistics 19:1, S. 1-24.

COSMAS II (2005): Corpus Search, Management and Analysis System unter www.ids-mannheim.de/cosmas2/, Stand: 17.11.2005.

DEREKO (2005): Deutsches Referenzkorpus unter www.ids-mannheim.de/projekte/korpora/, Stand: 17.11.2005.

Erben, Johannes (2003): mündlicher Beitrag auf der Tagung ‚Korpuslinguistik deutsch: synchron – diachron – kontrastiv'; 20.-23.3.2003, Universität Würzburg.

Frawley, William J. / Piatetsky-Shapiro, Gregory /Matheus, Christopher J. (1992): Knowledge Discovery in Databases: An Overview. In: AI Magazine**,** 13, S. 57-70.

Hanks, Patrick (2004): The Syntagmatics of Metaphor and Idiom. In: International Journal of Lexicography, 17, S. 245-274.

Köhler, Reinhard (2005): Korpuslinguistik – zu wissenschaftstheoretischen Grundlagen und methodologischen Perspektiven. In: LDV Forum, Band 20, Heft 2, S. 1-16.

Mukherjee, Joybrato (2002): Korpuslinguistik und Englischunterricht: Eine Einführung. Frankfurt / M.: Lang.

Perkuhn, Rainer / Belica, Cyril / al-Wadi, Doris / Lauer, Meike / Steyer, Kathrin / Weiß, Christian (2006): Korpustechnologie am Institut für Deutsche Sprache. In: Schwitalla, Johannes / Wegstein, Werner (Hrsg.): Korpuslinguistik deutsch: synchron – diachron – kontrastiv. Würzburger Kolloquium 2003, 20. – 23.3. 2003, Universität Würzburg. Tübingen: Niemeyer, erscheint 2006.

Sinclair, John (1991): Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Steyer, Kathrin (2004): Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven. In: Steyer, Kathrin (Hrsg.): Wortverbindungen – mehr oder weniger fest. Berlin/New York: de Gruyter (= Jahrbücher des Instituts für Deutsche Sprache 2003), S. 87-116.

Tognini-Bonelli, Elena (2001): Corpus Linguistics at Work. Amsterdam: Benjamins. (=Studies in Corpus linguistics 6)

Wittgenstein, Ludwig (1984): Philosophische Untersuchungen. Frankfurt / M.: Suhrkamp (erstmals veröffentlicht 1953).

Wittgenstein, Ludwig (1994): Tractatus Logico-Philosophicus. Frankfurt / M.: Suhrkamp (erstmals veröffentlicht 1922).

Anmerkungen

[1] Unter Data Mining versteht man das systematische (in der Regel automatisierte oder halbautomatische) Entdecken und Extrahieren unbekannter Informationen aus großen Mengen von Daten: „The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" (Frawley et al. 1992). Die so genannte Warenkorbanalyse ist der wohl klassischste Vertreter der Abhängigkeitsanalyse, bei der Informationen über das gleichzeitige Interesse für mehrere Leistungen oder Leistungsgruppen analysiert und in wirtschaftliches Verhalten umgesetzt werden.

[2] Jeder Beleg spiegelt ein singuläres Ereignis wider. Insofern ist fraglich, was ein „guter" Beleg ist. Vgl. auch Mythos 3.

[3] Abgesehen von dem Verzug durch die technische Aufbereitung ist die deskriptive Auslegung der Präsensformulierung des Mythos legitim. In diesem Abschnitt wird der präskriptiv gemeinte bzw. auf die Zukunft bezogene Gebrauch diskutiert.

[4] Im Sinne der Gebrauchstheorie Wittgensteins (Wittgenstein 1984), für den Teil des Sprachvermögens, der über die elementaren Möglichkeiten einer Abbildtheorie (Wittgenstein 1994) hinausgeht: „Es ist eine Hauptquelle unseres Unverständnisses, daß wir den Gebrauch unserer Wörter nicht übersehen" (Wittgenstein 1984, §122).

Die Autoren sind wissenschaftliche Mitarbeiter am Institut für Deutsche Sprache in Mannheim.